ALIGN, THEN SAMPLE: REFERENCE REGION ALIGNMENT FOR EQUIVARIANCE AND INVARIANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrary to the common assumption, convolutional neural networks (CNNs) are not inherently shift equivariant or invariant and a small shift in the input can lead to a drastic change in the output. Sampling operations, including downsampling and upsampling, are key components that can disrupt shift equivariance or invariance in CNN architectures. Conversely, sampling operations are important to improve computational efficiency and enlarge the receptive field for more contextual information. Furthermore, convolutional operations themselves break rotation and reflection invariance. Existing solutions typically address this issue at the model level rather than the individual layer level or rely on additional trainable parameters. In this paper, we introduce the reference region alignment (RRA) operation enhancing existing standard sampling methods (e.g., max pooling or average pooling), making them truly equivariant to arbitrary shifts, discrete 90° rotations and horizontal or vertical reflections. When integrated into CNN architectures, the RRA operation instills exact shift and roto-reflection equivariance at the individual layer level even before training, without introducing new trainable parameters. Experimental results demonstrate that integrating RRA with existing standard sampling operators improves CNN performance, yielding accuracy and consistency that are superior or comparable to established shift-equivariant and shift-invariant sampling methods across classification and semantic segmentation tasks, while also exhibiting roto-reflection consistency.

1 INTRODUCTION

Equivariance or invariance to different geometric affine transformations (e.g., translation, rotation, reflection and scale) is a highly useful property for any visual recognition model. The standard convolutional neural networks (CNNs) are neither equivariant nor invariant (Zhang, 2019; Cohen & Welling, 2016) to any geometric affine transformation. A small geometric change in the input can lead to a significant degradation in the network performance (Azulay & Weiss, 2019). As a result, a new class of CNNs has been developed that is equivariant or invariant to various geometric transformations (Cohen & Welling, 2016; Hoogeboom et al., 2018; Bekkers, 2019; Weiler & Cesa, 2019; Bekkers et al., 2018; Marcos et al., 2017; Rahman & Yeh, 2023). Earlier, it was assumed that fundamental layers of standard CNNs were shift-invariant, which means shifting the row or column of an input image has no effect on the output of a CNN. Thus, the primary focus of developing equivariant or invariant CNNs is either on rotation, scale, or reflection. However, the recent works (Azulay & Weiss, 2019; Zhang, 2019) have shown that the CNNs are not shift equivariant or invariant which is contrary to the prior assumptions. CNNs are not only sensitive to small shifts but their performance degrades substantially when the input is shifted by merely a single pixel (Azulay & Weiss, 2019). The feature learning block of CNNs mainly consists of arbitrary combinations of convolutional layers, pooling layers and activation functions. The convolution operation and activation functions are equivariant to shifts (Cohen & Welling, 2016).The sampling operations, essential in CNNs for enhancing computational efficiency and expanding the area of receptive field for more contextual information, whether achieved through pooling or strided convolutions, are one of the primary factors that disrupt shift-invariance in CNNs. This behavior arises because the fixed pooling windows sample different regions of the input after the shift (Zhang, 2019; Azulay & Weiss, 2019). Additionally, convolution operations in general are not equivariant to neither axis-aligned rotations nor reflections.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In the recent years, various efforts have been made to make CNNs shift invariant, ranging from straightforward solutions such as data augmentation (Fawzi et al., 2016), through blur-based downsampling (Gonzalez, 2009; Zhang, 2019) to adaptive subsampling grids (Chaman & Dokmanić, 2021; Rojas-Gomez et al., 2022). However, all of these techniques are conditioned to use the global pooling layer in the end to restore the changes caused by changes in the input to achieve shift-invariance or rely on additional learnable parameters.

In this paper, we present Reference Region Alignment (RRA), a parameter-free technique designed to enhance standard sampling operators by embedding equivariance directly into them. The proposed RRA method has several advantages over the state-of-the-art approaches (Azulay & Weiss, 2019; Zhang, 2019; Chaman & Dokmanić, 2021; Rojas-Gomez et al., 2022):

- guaranteed equivariance to arbitrary shifts, discrete axis-aligned rotations, and reflections;
- the first approach to introduce equivariance into existing sampling operations, maintaining equivariance within each individual layer without any additional trainable parameters;
- no reliance on global average pooling at the final layer to achieve invariance, instead operating seamlessly at every level;
- a general approach which can be used with other sampling operators (Scherer et al., 2010; Zeiler & Fergus, 2013; Lee et al., 2016).

Our experiments on CIFAR-10 (Krizhevsky et al., 2009), ImageNet (Deng et al., 2009), and PASCAL VOC (Everingham et al., 2015) demonstrate that RRA improves standard downsampling and upsampling operations, achieving accuracy and shift consistency that are superior or comparable to benchmark shift-equivariant and shift-invariant methods across image classification and semantic segmentation, while also maintaining rotoreflection consistency.

2 Related work

Embedding invariance or equivariance in neural network architectures is important because it enables higher consistency and accuracy, even when large datasets and sufficient compute power is available (Brehmer et al., 2024). The robustness of neural networks to essential geometric transformations such as shift or translation, rotation, scale, and reflection, has been deeply examined in numerous studies (Barnard & Casasent, 1990; Ilyas et al., 2019; Cohen & Welling, 2016; Azulay & Weiss, 2019; Hoogeboom et al., 2018). It is shown that neural networks do not exhibit invariance or equivariance to any of aforementioned geometric affine transformations.

Data augmentation (Cubuk et al., 2019; Fawzi et al., 2016; Hernández-García & König, 2018; Ho et al., 2019; Kauderer-Abrams, 2017) is a widely used and very simple method to improve the robustness of neural networks against random shifts and geometric transformations such as rotation, reflection and scaling. However, there are many downsides of this approach. First, data augmentation does not provide guaranteed invariance to out-of-distribution samples which were not encountered during training. Second, it captures geometric invariance only at global level. Lastly, data augmentation consumes useful network capacity in learning geometric features rather than focusing on task relevant representation learning (Lafarge et al., 2021; Brehmer et al., 2024).

As a result, a new class of neural networks has been designed with a focus to achieve invariance or equivariance to various type of geometric transformation and maintain consistent performance across these transformations (Cohen & Welling, 2016; Chaman & Dokmanić, 2021; Bekkers, 2019; Weiler & Cesa, 2019; Worrall et al., 2017). Canonicalization-based methods achieve invariance or equivariance by mapping transformed inputs to a canonical representation before prediction (Kaba et al., 2023); unlike these learned approaches, RRA performs deterministic, parameter-free canonical alignment. Anti-aliasing (Azulay & Weiss, 2019) strategies applied to intermediate feature representations have shown potential to improve shift robustness, but these methods offer only partial solutions. One approach to address this challenge involves modifying the max pooling operation. Instead of applying max pooling followed directly by subsampling, the operation is reformulated as a two-step process comprising a dense max operation followed by naive subsampling (Zhang, 2019). A low-pass filter (LPF) is inserted between these two steps to reduce aliasing. Next, a content-aware low-pass filtering mechanism has been proposed (Zou et al., 2023) which dynamically predicts filters for each spatial location based on the local frequency content of feature maps. However, blur-based

downsampling in CNNs improve the network performance but the non-linear activation function amplifies the small differences and a beyond certain point, over-blurring reduces the network performance. Another technique, known as Adaptive Polyphase Sampling (APS) (Chaman & Dokmanić, 2021), achieves improved shift invariance by adaptively selecting subsampling grids based on high-energy regions, as opposed to using fixed grid patterns. Learnable polyphase sampling (LPS) (Rojas-Gomez et al., 2022), a generalization of APS, further advances this concept by introducing a pair of trainable downsampling and upsampling layers. These components are designed to achieve true shift equivariance, offering better generalization to unseen spatial transformations.

3 METHODOLOGY

3.1 EQUIVARIANCE AND INVARIANCE

A symmetry group is a tuple $(G, \cdot)$ consisting of a set G and a binary operation $\cdot : G \times G \to G, (g, h) \mapsto g \cdot h$, which satisfy the group axioms of *associativity, identity* and *inverses*. Further, let X be a set. A (left) group action is a map $\triangleright : G \times X \to X, (g, x) \mapsto g \triangleright x$, compatible with group composition and identity axioms.

Let $\triangleright^X$ and $\triangleright^Y$ be group actions of a group G on sets X and Y. A function (or an operator) $L : X \to Y$ is G-**equivariant** *iff* it satisfies:

$$L(g \triangleright^X x) = g \triangleright^Y L(x) \quad \forall g \in G, x \in X. \tag{1}$$

Invariance is a special case of equivariance, where a function $L : X \to Y$ is considered G-**invariant** *iff* the group action $\triangleright^Y$ on the codomain is a trivial action id_Y:

$$L(g \triangleright^X x) = g \, \mathrm{id}_Y L(x) = L(x) \quad \forall g \in G, x \in X. \tag{2}$$

To apply this to discrete 2D images, let $\mathbf{F} \in \mathcal{F}_{N,C}$ be a square multichannel image or tensor $\mathbf{F} : \mathbb{Z}_N^2 \times \mathbb{Z}_C \to \mathbb{Z}_{2^B}$ (where $\mathbb{Z}_N = \{0, \ldots, N-1\}$, and B is the bit-depth of the image). $\mathbf{F}_c$ indexes the channel $c \in \mathbb{Z}_C$, and $\mathbf{F}[\mathbf{x}]$ indexes the spatial position $\mathbf{x} \in \mathbb{Z}_N^2$. The circular shift group action of the discrete translation group $(\mathbb{Z}^2, +)$ on $\mathcal{F}^{N,C}$ is defined as:

$$(\mathbf{t} \triangleright_{N,C}^T \mathbf{F})_c[\mathbf{x}] = \mathbf{F}_c[\mathbf{x} - \mathbf{t} \pmod N] \quad \forall \mathbf{t} \in \mathbb{Z}^2, \mathbf{x} \in \mathbb{Z}_N^2. \tag{3}$$

The definition of **shift-equivariance** and **invariance** is obtained by substituting $G = (\mathbb{Z}^2, +), X = \mathcal{F}_{N,C}$ in Equation 1 and Equation 2. Global pooling operators, such as global average pooling, are shift-invariant.

The group action of the dihedral group $D_4 = \langle r, s | r^4, s^2, (sr)^2 \rangle$ (where r is a counterclockwise rotation for $\frac{\pi}{2}$, and s is a reflection about the horizontal axis) on $\mathcal{F}^{N,C}$ is given by:

$$(g \triangleright_{N,C}^D \mathbf{F})_c[\mathbf{x}] = \mathbf{F}_c[g^{-1}(\mathbf{x})] \quad \forall g \in D_4, \mathbf{x} \in \mathbb{Z}_N^2. \tag{4}$$

The definition of D_4-equivariance and invariance is obtained by substituting $G = D_4, X = \mathcal{F}_{N,C}$ in Equation 1 and Equation 2. Since the dihedral group is a semidirect product of the cyclic group C_4 containing rotations by multiples of $\frac{\pi}{2}$, and the group $(\{\pm 1\}, *)$ containing reflections along a given axis, $D_4 \cong C_4 \rtimes (\{\pm 1\}, *)$, D_4-equivariant (or -invariant) operators are **rotation and reflection equivariant** (or **invariant**).

To achieve equivariance to shift, rotation and reflection, we consider the group $(\mathbb{Z}^2, +) \rtimes D_4$. The group action on $\mathcal{F}_{N,C}$ is given by:

$$\begin{aligned}(\mathbf{t}g \triangleright_{N,C} \mathbf{F})_c[\mathbf{x}] &= \mathbf{t} \triangleright_{N,C}^T (g \triangleright_{N,C}^D \mathbf{F})_c[\mathbf{x}] \\ &= \mathbf{F}_c[g^{-1}(\mathbf{x} - \mathbf{t}) \pmod N], \quad \forall \mathbf{t} \in \mathbb{Z}^2, g \in D_4, \mathbf{x} \in \mathbb{Z}_N^2.\end{aligned} \tag{5}$$

Since this group is a semidirect product of $(\mathbb{Z}^2, +)$ and D_4, the following property holds:

$$g \triangleright_{N,C}^D \mathbf{t} \triangleright_{N,C}^T \mathbf{F} = g\mathbf{t} \triangleright_{N,C}^T g \triangleright_{N,C}^D \mathbf{F} \tag{6}$$

In the next subsection, we present the RRA operator which is **shift, rotation and reflection-equivariant**.

3.2 REFERENCE REGION ALIGNMENT (RRA)

Consider a single-channel image $F \in \mathcal{F}_{N,1}$. To identify a **reference region** of order k, consider a candidate neighbourhood of size $k \times k$ for each spatial location $(i, j) \in \mathbb{Z}_N^2$, using circular padding:

$$Q^k_{(i,j)}[x, y] = F[x + i \pmod N, y + j \pmod N] \quad \forall (x, y) \in \mathbb{Z}_k^2. \tag{7}$$

For each candidate region $Q^k_{(i,j)}$, compute directional signature R, C, T capturing spatial symmetry, and a helper signature T':

$$R^k_{(i,j)} = \sum_{x=0}^{\lfloor \frac{k}{2}-1 \rfloor} \sum_{y=0}^{k-1} Q^k_{(i,j)}[x, y] - \sum_{x=\lceil \frac{k}{2} \rceil}^{k-1} \sum_{y=0}^{k-1} Q^k_{(i,j)}[x, y] \tag{8}$$

$$C^k_{(i,j)} = \sum_{x=0}^{k-1} \sum_{y=0}^{\lfloor \frac{k}{2}-1 \rfloor} Q^k_{(i,j)}[x, y] - \sum_{x=0}^{k-1} \sum_{y=\lceil \frac{k}{2} \rceil}^{k-1} Q^k_{(i,j)}[x, y] \tag{9}$$

$$T^k_{(i,j)} = \sum_{x=0}^{k-1} \sum_{y>x} Q^k_{(i,j)}[x, y] - \sum_{x=0}^{k-1} \sum_{y<x} Q^k_{(i,j)}[x, y] \tag{10}$$

$$T'^k_{(i,j)} = \sum_{x=0}^{k-1} \sum_{y<k-1-x} Q^k_{(i,j)}[x, y] - \sum_{x=0}^{k-1} \sum_{y>k-1-x} Q^k_{(i,j)}[x, y] \tag{11}$$

Order k is assumed for the remainder of this section, so the order index will be omitted for brevity.

A score is then calculated for each candidate region $Q_{(i,j)}$ and each $g \in D_4$, e.g. as the sum of directional signatures $S(g \triangleright^D Q) = g \triangleright^D R + g \triangleright^D C + g \triangleright^D T$, or their lexicographical ordering. The effect of the group action of D_4 on the directional signatures is shown in Table 1. The **reference region** is defined as the candidate region with the highest score for any $g \in D_4$, and specified by its location $(i^r, j^r) \in \mathbb{Z}_N^2$:

$$(i^r, j^r) = \arg \max_{(i,j) \in \mathbb{Z}_N^2} S(g \triangleright^D Q_{i,j}) \quad \forall g \in D_4. \tag{12}$$

Table 1: The first three rows show the effect of the group action of D_4 on the directional signatures $R = R^k_{(i,j)}, C = C^k_{(i,j)}, T = T^k_{(i,j)}$ of a candidate region $Q^k_{(i,j)}$ for each $g \in D_4$. The last two rows show how the coordinates of the upper-left corner of a 2×2 square transform under g.

$\mathbf{g} \in \mathbf{D_4}$	id	$\mathbf{r}$	$\mathbf{r^2}$	$\mathbf{r^3}$	$\mathbf{s}$	$\mathbf{rs}$	$\mathbf{r^2s}$	$\mathbf{r^3s}$
$\mathbf{R}$	R	$-C$	$-R$	C	$-R$	$-C$	R	C
$\mathbf{C}$	C	R	$-C$	$-R$	C	$-R$	$-C$	R
$\mathbf{T}$	T	$-T'$	$-T$	T'	$-T'$	T	T'	$-T$
orow	0	1	1	0	1	1	0	0
ocol	0	0	1	1	0	1	1	0

The orientation $g^* \in D_4$ for which the maximum score of the reference region is achieved is then considered the **reference orientation**:

$$g^* = \arg \max_{g \in D_4} S(g \triangleright^D Q_{(i^r, j^r)}). \tag{13}$$

According to Table 1, if at least some of $R, C, T, T' \neq 0$, the maximum score at each location (i, j) is achieved for a unique $g \in D_4$ and corresponds to the transformation under which $R, C, T \geq 0$.

The **reference location** $\mathbf{t}^* = (i^*, j^*) \in \mathbb{Z}_N^2$ is defined as the coordinates of the upper-left corner of the reference region when the image is transformed to the reference orientation $g^* \triangleright^D F$. To obtain the reference location, we first take the original image coordinates $\mathbf{t'}^* = (i'^*, j'^*) \in \mathbb{Z}_N^2$ of the corner of the reference region which corresponds to its upper-left corner in its reference orientation:

$$\mathbf{t'}^* = (i'^*, j'^*) = (i^r + \mathrm{orow}(g^*)(k-1) \pmod N, j^r + \mathrm{ocol}(g^*)(k-1) \pmod N), \tag{14}$$

with $(\mathrm{ocol}(g), \mathrm{orow}(g))$ given in Table 1. Then, the reference location is obtained by applying the action g^* to $\mathbf{t'}^*$:

$$\mathbf{t}^* = g^*(\mathbf{t'}^*) \tag{15}$$

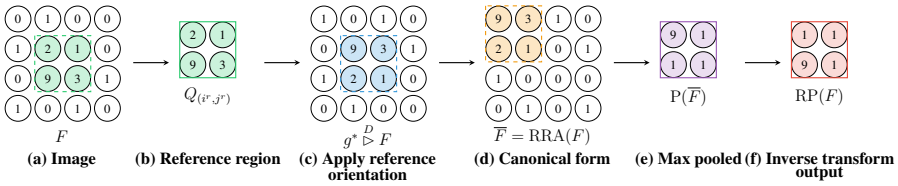

Figure 1: Illustration of the reference region alignment (RRA) operator combined with max pooling.

in the reference orientation coordinates.

Finally, we define a **reference region alignment** operator, $\mathrm{RRA}(F)$ (illustrated in Fig. 1) as the action of $-\mathbf{t}^*$ to the region in the reference orientation $g^* \triangleright^D F$:

$$\overline{F} = \mathrm{RRA}(F) = -\mathbf{t}^* \triangleright^T (g^* \triangleright^D F) = -\mathbf{t}^* g^* \triangleright F, \tag{16}$$

where $\overline{F}$ is the **canonical form of the image** F. The inverse operation is given by:

$$\mathrm{RRA}^{-1}(\overline{F}) = g^{*-1} \triangleright^D \mathbf{t}^* \triangleright^T \overline{F}. \tag{17}$$

Since a unique reference region, corresponding to the unique pair $(\mathbf{t}^*, g^*)$ can be found in any image given sufficiently large k, the inverse operation allows for a reconstruction of the original image F.

To extend the proposed RRA framework to multi-channel images such as the ones used in CNNs, $\mathbf{F} \in \mathcal{F}_{N,C}$, construct a 2D spatial projection of the image by aggregating responses across channels:

$$Y[\mathbf{x}] = \sum_{c=0}^{C-1} \mathbf{F}_c[\mathbf{x}] \quad \forall \mathbf{x} \in \mathbb{Z}_N^2, \tag{18}$$

which captures the overall activation at each spatial location. The reference orientation and location g^* and $\mathbf{t}^*$ are obtained from Y, while the full image $\mathbf{F}$ is transformed according to Equation 16 and Equation 17.

3.3 Properties of RRA

In this section, we show the invariance of the $\mathrm{RRA}(\cdot)$ operator to $(\mathbb{Z}^2, +) \rtimes D_4$, as well as how equivariance can be achieved when combined with its inverse $\mathrm{RRA}^{-1}(\cdot)$.

Take an image $\mathbf{F}$ and its translated, rotated and reflected version, acted upon by some $\mathbf{t}g \in (\mathbb{Z}^2, +) \rtimes D_4$ and given by $\mathbf{t}g \triangleright \mathbf{F} = \mathbf{t} \triangleright^T g \triangleright^D \mathbf{F}$.

Express the reference location and orientation of the transformed image using Equation 13, Equation 14 and Equation 15 in terms of the reference location and orientation in the original image:

$$g^*(\mathbf{t}g \triangleright \mathbf{F}) = g^* g^{-1} \tag{19}$$

$$\mathbf{t}'^*(\mathbf{t}g \triangleright \mathbf{F}) = g\mathbf{t}'^* + \mathbf{t} \tag{20}$$

$$\mathbf{t}^*(\mathbf{t}g \triangleright \mathbf{F}) = g^* g^{-1}(g\mathbf{t}'^* + \mathbf{t}). \tag{21}$$

To obtain $\mathrm{RRA}(\mathbf{t}g \triangleright \mathbf{F})$, substitute Equation 19, Equation 20 and Equation 21 into Equation 16:

$$\mathrm{RRA}(\mathbf{t}g \triangleright \mathbf{F}) = -g^* g^{-1}(g\mathbf{t}'^* + \mathbf{t}) \triangleright^T g^* g^{-1} \triangleright^D \mathbf{t}g \triangleright \mathbf{F}. \tag{22}$$

From Equation 22), by using the property in Equation 6:

$$\begin{aligned}
\mathrm{RRA}(\mathbf{t}g \triangleright \mathbf{F}) &= -g^* g^{-1}(g\mathbf{t}'^* + \mathbf{t}) \triangleright^T g^* g^{-1} \triangleright^D \mathbf{t}g \triangleright \mathbf{F} \\
&= g^* g^{-1} \triangleright^D -(g\mathbf{t}'^* + \mathbf{t}) \triangleright^T \mathbf{t}g \triangleright \mathbf{F} \\
&= g^* g^{-1} \triangleright^D -g\mathbf{t}'^* \triangleright^T g \triangleright^D \mathbf{F} \\
&= g^* g^{-1} \triangleright^D -gg^{*-1}\mathbf{t}^* \triangleright^T g \triangleright^D \mathbf{F} \\
&= -\mathbf{t}^* \triangleright^T g^* g^{-1} \triangleright^D g \triangleright^D \mathbf{F} \\
&= -\mathbf{t}^* \triangleright^T g^* \triangleright^D \mathbf{F} = \overline{\mathbf{F}} = \mathrm{RRA}(\mathbf{F}),
\end{aligned} \tag{23}$$

we show that the $\mathrm{RRA}(\cdot)$ operator is invariant to $(\mathbb{Z}^2, +) \rtimes D_4$, and is therefore shift, rotation and reflection invariant.

Further, consider the max pooling operator of size 2×2 and with stride 2, $\mathrm{P} : \mathcal{F}_{N,C} \rightarrow \mathcal{F}_{\lfloor N/2 \rfloor, C}$:

$$\mathrm{P}(\mathbf{F})_c[x, y] = \max(\mathbf{F}[2x, 2y], \mathbf{F}[2x+1, 2y], \mathbf{F}[2x, 2y+1], \mathbf{F}[2x+1, 2y+1]). \tag{24}$$

For $g \in D_4$, the same definition of group action holds on $\mathcal{F}_{N,C}$ and $\mathcal{F}_{\lfloor N/2 \rfloor, C}$. To show equivariance to $(\mathbb{Z}^2, +) \rtimes D_4$, we define the group action of $(\mathbb{Z}^2, +)$ on $\mathcal{F}_{\lfloor N/2 \rfloor, C}$ as:

$$(\mathbf{t} \rhd^T_{\lfloor N/2 \rfloor, C} \mathbf{F})_c[\mathbf{x}] = \mathbf{F}_c[\mathbf{x} - \lfloor \frac{\mathbf{t}}{2} \rfloor \pmod{\lfloor \frac{N}{2} \rfloor}] \quad \forall \mathbf{t} \in \mathbb{Z}^2, \mathbf{x} \in \mathbb{Z}^2_{\lfloor N/2 \rfloor}. \tag{25}$$

A new sampling operator $\mathrm{RP} : \mathcal{F}_{N,C} \rightarrow \mathcal{F}_{\lfloor N/2 \rfloor, C}$ is defined by combining $\mathrm{P}(\cdot)$ with $\mathrm{RRA}(\cdot)$:

$$\mathrm{RP}(\mathbf{F}) = \mathrm{RRA}^{-1}(\mathrm{P}(\mathrm{RRA}(\mathbf{F}))). \tag{26}$$

It can be shown that such a sampling operator is equivariant to $(\mathbb{Z}^2, +) \rtimes D_4$ according to equation 27. Specifically, it can be obtained:

$$\begin{aligned}
\mathrm{RP}(\mathbf{t}g \rhd_N \mathbf{F}) &= \mathbf{t}g \rhd_{\lfloor \frac{N}{2} \rfloor} \mathrm{RP}(\mathbf{F}) \\
\mathrm{RRA}^{-1}(\mathrm{P}(\mathrm{RRA}(\mathbf{t}g \rhd_N \mathbf{F}))) &= \mathbf{t}g \rhd_{\lfloor \frac{N}{2} \rfloor} \mathrm{RRA}^{-1}(\mathrm{P}(\mathrm{RRA}(\mathbf{F}))) \\
g^{*-1}g \rhd^D_{\lfloor \frac{N}{2} \rfloor} g^* g^{-1}(g\mathbf{t}'^* + \mathbf{t}) \rhd^T_{\lfloor \frac{N}{2} \rfloor} \mathrm{P}(\overline{\mathbf{F}}) &= \mathbf{t} \rhd^T_{\lfloor \frac{N}{2} \rfloor} g \rhd^D_{\lfloor \frac{N}{2} \rfloor} g^{*-1} \rhd^D_{\lfloor \frac{N}{2} \rfloor} \mathbf{t}^* \rhd^T_{\lfloor \frac{N}{2} \rfloor} \mathrm{P}(\overline{\mathbf{F}}) \\
\mathbf{t} \rhd^T_{\lfloor \frac{N}{2} \rfloor} gg^{*-1}\mathbf{t}^* \rhd^T_{\lfloor \frac{N}{2} \rfloor} g^{*-1}g \rhd^D_{\lfloor \frac{N}{2} \rfloor} \mathrm{P}(\overline{\mathbf{F}}) &= \mathbf{t} \rhd^T_{\lfloor \frac{N}{2} \rfloor} gg^{*-1} \rhd^D_{\lfloor \frac{N}{2} \rfloor} \mathbf{t}^* \rhd^T_{\lfloor \frac{N}{2} \rfloor} \mathrm{P}(\overline{\mathbf{F}}) \\
\mathbf{t} \rhd^T_{\lfloor \frac{N}{2} \rfloor} gg^{*-1} \rhd^D_{\lfloor \frac{N}{2} \rfloor} \mathbf{t}^* \rhd^T_{\lfloor \frac{N}{2} \rfloor} \mathrm{P}(\overline{\mathbf{F}}) &= \mathbf{t} \rhd^T_{\lfloor \frac{N}{2} \rfloor} gg^{*-1} \rhd^D_{\lfloor \frac{N}{2} \rfloor} \mathbf{t}^* \rhd^T_{\lfloor \frac{N}{2} \rfloor} \mathrm{P}(\overline{\mathbf{F}}).
\end{aligned} \tag{27}$$

Therefore, the operator constructed by combining $\mathrm{RRA}(\cdot)$ and a sampling operation (such as max pooling) is equivariant to $(\mathbb{Z}^2, +) \rtimes D_4$. Note also that the proof of equivariance in Equation 27 is independent of the choice of sampling operation given in Equation 24, and holds for any operation as long as the group actions of D_4 and $(\mathbb{Z}^2, +)$ are defined of the operators codomain. Therefore, $\mathrm{RRA}(\cdot)$ can be also be combined with individual convolution operation, or the whole network, to achieve end-to-end invariance or equivariance.

4 Experiments and results

We evaluate the proposed RRA technique on image classification and semantic segmentation. Image classification experiments are conducted on CIFAR-10 (Krizhevsky et al., 2009) and ImageNet (Deng et al., 2009) under circular and standard shift settings, following established protocols (Chaman & Dokmanić, 2021; Rojas-Gomez et al., 2022). Semantic segmentation is evaluated on PASCAL VOC (Everingham et al., 2015) under circular and standard shifts, following prior segmentation consistency settings (Rojas-Gomez et al., 2022; Zou et al., 2020). We also report roto-reflection consistency for all experiments.

RRA is compared with conventional sampling methods (downsampling/upsampling), anti-aliased low-pass filtering (LPF) (Zhang, 2019), adaptive anti-aliasing using DDAC (Zou et al., 2020; 2023), Adaptive Polyphase Sampling (APS) (Chaman & Dokmanić, 2021), and Learnable Polyphase Sampling (LPS) (Rojas-Gomez et al., 2022), comprising Learnable Polyphase Downsampling (LPD) and Learnable Polyphase Upsampling (LPU). Experiments are performed using ResNet-based architectures (He et al., 2016), including ResNet-18, ResNet-50, and ResNet-101, where standard downsampling layers are replaced by the evaluated methods. Unless otherwise stated, RRA is implemented with $k = 3$ and uses max pooling as the base downsampling operator. We report accuracy and consistency metrics for both classification and semantic segmentation. In all experiments reporting roto-reflection consistency, RRA is also applied to the network input. All experiments involving the proposed RRA method are repeated 5 times, and the mean performance is reported.

4.1 Classification (Circular Shift + Roto-reflection)

Experiment details. We evaluate classification accuracy and transformation consistency on CIFAR-10 and ImageNet using ResNet-18. Following the circular-shift setting of Chaman & Dokmanić

Table 2: Circular consistency and top-1 accuracy on CIFAR-10 and ImageNet using ResNet-18 with different downsampling methods. The anti-aliasing setting is reported where applicable.

Method	Anti-Alias	CIFAR-10		ImageNet	
		C-Cons. ↑	Accuracy ↑	C-Cons. ↑	Accuracy ↑
Baseline	–	89.64	91.44	80.39	64.88
LPF	Bin-5	98.34	94.43	84.37	67.05
DDAC	Adapt-3	95.13	93.17	77.23	67.60
APS	Adapt-3	100.00	94.43	100.00	69.02
LPD	Adapt-3	100.00	94.57	100.00	69.11
RRA (Ours)	Adapt-3	**100.00**	**94.60**	**100.00**	**69.26**

(2021), CIFAR-10 models are trained for 250 epochs with a batch size of 256 using stochastic gradient descent (SGD), momentum of 0.9, weight decay of 5×10^{-4}, and an initial learning rate of 0.1 reduced by a factor of 0.1 every 100 epochs. ImageNet models are trained with SGD for 90 epochs with a batch size of 64, momentum of 0.9, weight decay of 10^{-4}, and an initial learning rate of 0.1 reduced by a factor of 0.1 every 30 epochs. For CIFAR-10, the original 32×32 images are normalized using the training set normalisation statistics, with horizontal flipping ($p = 0.5$) during training. For ImageNet, images are resized such that the shorter side is 256 pixels and center-cropped to 224×224, with random horizontal flipping during training and normalization using ImageNet statistics. We compare the proposed RRA with DDAC, APS and LPD (all using a 3×3 adaptive anti-aliasing filter, Adapt-3), and with LPF (which uses a Bin-5).

Evaluation metrics. Top-1 classification accuracy is the percentage of correctly predicted labels for the original, unshifted test set. Circular consistency (C-Cons.) follows the established circular-shift evaluation protocols (Chaman & Dokmanić, 2021; Rojas-Gomez et al., 2022), and is defined as the average percentage of predicted labels that are equal under two different circular shifts, i.e.:

$$T_\theta \left(\mathbf{t}_1 \triangleright^T \mathbf{F} \right) = T_\theta \left(\mathbf{t}_2 \triangleright^T \mathbf{F} \right), \tag{28}$$

where $\mathbf{F}$ is an input image, $\mathbf{t}_1 = (i_1, j_1)$ and $\mathbf{t}_2 = (i_2, j_2)$ denote two circular translations sampled independently and uniformly at random from $\{0, \ldots, 31\}^2$, and T_θ is the trained CNN model. Similarly, to evaluate roto-reflection consistency (D_4-Cons.), we compare the predicted class labels of two independently transformed versions of the same input image under elements of D_4:

$$T_\theta \left(g_1 \triangleright^D \mathbf{F} \right) = T_\theta \left(g_2 \triangleright^D \mathbf{F} \right), \tag{29}$$

where $g_1, g_2 \in D_4$ denote two roto-reflection transformations sampled independently and uniformly at random from the dihedral group D_4.

CIFAR-10 and ImageNet results. Table 2 reports classification consistency and top-1 accuracy on CIFAR-10 and ImageNet. Across both datasets, RRA achieves 100% C-Cons., matching APS and LPD while improving over the baseline, LPF, and DDAC.On CIFAR-10, RRA obtains the highest accuracy of $94.60 \pm 0.08\%$, slightly outperforming LPD, APS, and LPF. On ImageNet, RRA also achieves the highest top-1 accuracy of $69.26 \pm 0.11\%$, improving over LPD, APS, DDAC, LPF, and the baseline while preserving perfect C-Cons. On both CIFAR-10 and ImageNet, RRA achieves 100% D_4-Cons. while maintaining the same accuracy. These results demonstrate that the proposed RRA downsampling strategy maintains exact transformation consistency without compromising classification accuracy across both CIFAR-10 and ImageNet.

4.2 Classification (Standard Shift + Roto-reflection)

Experiment details. We further evaluate classification accuracy and consistency under standard image shifts on ImageNet. Following the protocol used in prior work (Zhang, 2019; Rojas-Gomez et al., 2022), we compare with LPF and LPD using both ResNet-50 and ResNet-101, and with DDAC using ResNet-101 only. Training images are resized to 256×256 and augmented using resized random cropping to 224×224 and random horizontal flipping, with normalization with ImageNet statistics applied to both training and test images. Models are trained for 90 epochs using SGD with momentum 0.9, weight decay 10^{-4}, and an initial learning rate of 0.1, which is reduced by a factor of 0.1 every 30 epochs.

Table 3: Top-1 classification accuracy and standard shift consistency on ImageNet using ResNet-50/101 with different downsampling methods.

Method	ResNet-50			ResNet-101		
	Anti-Alias	Acc. ↑	S-Cons. ↑	Anti-Alias	Acc. ↑	S-Cons. ↑
Baseline	–	76.16	89.20	–	77.70	90.60
LPF	Tri-3	76.83	90.91	Tri-3	78.40	91.60
LPD	Tri-3	77.14	91.43	Adapt-3	78.80	92.40
DDAC	–	–	–	Adapt-3	**79.00**	91.80
RRA (Ours)	Tri-3	**77.21**	**91.89**	Adapt-3	78.95	**93.54**

Evaluation metrics. In this setting, zero-padding is used on the image boundaries, and the center region of size 224×224 is cropped for evaluation. Shift consistency (S-Cons.) follows the definition in Equation 28, except that circular shift is replaced with standard shift with zero-padding. Since standard shifts introduce boundary effects, perfect shift consistency is not expected. APS is omitted because the standard-shift setting follows the LPF/DDAC evaluation protocol. We also evaluate D_4-consistency using Equation 29.

Results. Table 3 reports top-1 accuracy and S-Cons. on ImageNet using ResNet-50 and ResNet-101. For ResNet-50, RRA achieves the highest accuracy of $77.21 \pm 0.05\%$ and S-cons. of $91.89 \pm 0.18\%$, improving over the baseline, LPF, and LPD. For ResNet-101, RRA obtains the highest S-Cons. of $93.54 \pm 0.12\%$ while maintaining accuracy comparable to DDAC, whereas DDAC achieves the highest accuracy. With both settings, RRA maintains 100% D_4-Cons. These results show that RRA remains effective under the more challenging standard-shift setting, where the theoretical circular-shift assumptions are no longer exactly satisfied.

4.3 Semantic Segmentation

Experiment details. We evaluate RRA on the PASCAL VOC dataset for semantic segmentation under both circular and standard shift settings. Following LPS (Rojas-Gomez et al., 2022), we use DeepLabV3+ (Chen et al., 2018) with a ResNet-18 backbone for circular-shift and a ResNet-101 backbone for standard-shift, both initialized from ImageNet-pretrained models. For semantic segmentation, RRA is also applied to upsampling operations; when evaluating roto-reflection consistency, the inverse RRA is applied to the predicted segmentation map. This preserves the spatial correspondence between the input image and the predicted mask. We compare RRA with DeepLabV3+, LPF, DDAC, and LPS (LPD+LPU). Training images are randomly scaled by a factor in $[0.5, 2.0]$, cropped to 513×513, randomly horizontally flipped, and normalized using ImageNet statistics; validation images are resized to 513 pixels, center-cropped to 513×513, and normalized using the same statistics. Following the LPS segmentation protocol, models are trained for 125 epochs. Optimization is performed using SGD with momentum 0.9, weight decay 10^{-4}, an initial learning rate of 0.01, and a polynomial learning-rate schedule, with cross-entropy loss as the training objective.

Evaluation metrics. Segmentation performance is reported as mean intersection over union (mIoU) in case of both circular and standard shifts. Shift-equivariance under circular shifts is evaluated using mean Average Segmentation Circular Consistency (mASCC). Predictions obtained from shifted inputs are shifted back to the original spatial coordinates before their pixel-wise consistency is calculated. For standard shifts, equivariance is evaluated using mean Average Semantic Segmentation Consistency (mASSC), which is the linear-shift counterpart of mASCC. Boundary pixels affected by shifting and cropping are excluded from the calculation. For roto-reflections, we report mean Average Semantic Rotoreflection Consistency (mASRC), the roto-reflection counterpart of the metric.

Circular shift results. As shown in Table 4 (left), RRA achieves the highest reported mIoU of 72.47 ± 0.07 while attaining an mASCC of 100, an improvement of 0.05% in mIoU over LPS. RRA also achieves mASRC of 100.

Standard shift results. As shown in Table 4 (right), RRA obtains the highest reported values, achieving an mIoU of 80.45 ± 0.05 and an mASSC of 98.50 ± 0.13. Compared with LPS, RRA improves mIoU by 0.02% and mASSC by 0.52%. RRA also obtains mASRC of 100, improving

Table 4: Semantic segmentation results on PASCAL VOC using DeepLabV3+ under different backbone and shift settings.

Method	ResNet-18, Circular shift			ResNet-101, Standard shift		
	Anti-Alias	mIoU	mASCC	Anti-Alias	mIoU	mASSC
DeepLabV3+	–	70.03	95.42	–	78.50	95.50
LPF	Rect-2	71.02	96.03	Tri-3	79.40	95.90
DDAC	Adapt-3	72.28	95.98	Adapt-3	80.30	96.30
APS	Adapt-3	72.37	96.70	–	–	–
LPS	Adapt-3	72.37	100.00	Adapt-3	80.43	97.98
RRA (Ours)	Adapt-3	**72.47**	**100.00**	Adapt-3	**80.45**	**98.50**

shift consistency under both settings and achieving perfect axis-aligned roto-reflection consistency, while maintaining competitive segmentation accuracy.

4.4 Time Analysis

Experiment details. We compare the proposed RRA with the max-pooling baseline, LPF, DDAC, APS, and LPD in terms of computational time and the requirement for learnable parameters. The experiment is conducted at the sampling-operator level. Runtime is measured on input feature maps of size $32 \times 64 \times 112 \times 112$, randomly generated from a standard normal distribution. For each operator, we perform 10 independent trials with 200 forward passes per trial after 50 warm-up iterations. All experiments are conducted on a single NVIDIA Tesla T4 GPU.

Table 5: Time comparison of downsampling methods, including the effect of reference-region size k on RRA. Results are reported as mean ($\pm$) standard deviation over repeated forward-pass measurements.

Method	Learnable params	Latency (ms)	
		Mean ↓	Std
Baseline (max pool)	No	0.90	0.05
LPF (Tri-3)	No	3.90	0.10
DDAC (Adapt-3)	Yes	27.51	0.76
APS + LPF (Tri-3)	No	7.30	0.15
APS + DDAC (Adapt-3)	Yes	30.85	0.90
LPD + LPF (Tri-3)	Yes	27.70	0.97
LPD + DDAC (Adapt-3)	Yes	51.69	1.53
RRA + max pool ($k = 3$)	No	8.18	0.15
RRA + max pool ($k = 5$)	No	8.85	0.23
RRA + max pool ($k = 7$)	No	9.18	0.13
RRA + LPF (Tri-3) ($k = 3$)	No	10.18	0.16
RRA + LPF (Tri-3) ($k = 5$)	No	10.89	0.34
RRA + DDAC (Adapt-3) ($k = 3$)	Yes	33.79	0.48
RRA + DDAC (Adapt-3) ($k = 5$)	Yes	34.53	0.67

Results. Table 5 compares the computational efficiency of the evaluated downsampling methods. The baseline (max pool) and LPF incur the lowest computational cost, while DDAC- and LPD-based variants are more expensive. RRA requires no learnable parameters and achieves 8.18 ± 0.15 ms with $k = 3$. RRA is considerably faster than DDAC and LPD and comparable to APS.

5 Conclusion

We propose Reference Region Alignment (RRA), a general alignment technique that embeds equivariance directly into standard downsampling and upsampling layers. When integrated into CNN architectures, RRA ensures exact equivariance under circular shifts, 90° rotations, and horizontal/vertical reflections. RRA is not restricted to a specific pooling layers and can be combined with different down/upsampling layers, including max pooling, and anti-aliased LPF. Experimental results on CIFAR-10, ImageNet, and PASCAL VOC demonstrate that RRA achieves superior or comparable performance in classification accuracy, transformation consistency, and semantic segmentation consistency under both circular-shift and standard-shift evaluation settings.

AI USE STATEMENT

In this work, we used generative AI tools only to edit the research paper to improve readability, table formatting, and for TikZ illustration generation. We have not used generative AI tools to develop theoretical models or conceptual frameworks, formulate mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design research methodology, implement methods, support data analysis or interpret results. The tasks of generating synthetic data, translation and cleaning or reformatting datasets are not applicable to this work. We have reviewed all AI-assisted work by verifying that the language edits preserved the intended technical meaning and scientific claims, the content of generated TikZ illustrations is the same as that of our hand-drawn sketches, and by verifying that all of the numbers reported in the reformatted tables match the ones directly obtained in our experiments. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The research presented herein adheres to the ICLR Code of Ethics. All experiments are performed using publicly available datasets (CIFAR-10, ImageNet, PASCAL VOC), and do not involve human subjects, personal or otherwise sensitive information. To the best of the authors' knowledge, this research does not pose any significant ethical or societal risks beyond those generally associated with the development and use of machine learning methods. The authors declare that they have no financial or non-financial conflicts of interest and that this work received no sponsorship that could have influenced the research reported herein.

REPRODUCIBILITY STATEMENT

The proposed method is described in full in Sec. 3.2, and the complete proof of the both main claims in the paper is included in Sec. 3.3 (in Equation 23 and Equation 27). An anonymous downloadable source code has been submitted as supplementary materials, and will be shared as open-source under a non commercial license upon acceptance. The code contains all data processing steps, hyperparameters and experimental settings (described in full in Sec. 4), and the implementation of the proposed method.

AUTHOR CONTRIBUTIONS

J. S.: Conceptualization, Methodology, Software, Validation, Investigation, Writing – Original Draft, Writing – Review & Editing, Visualization. **P. B.**: Conceptualization, Methodology, Formal Analysis, Investigation, Writing – Original Draft, Writing – Review & Editing.

REFERENCES

Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research*, 20(184):1–25, 2019.

Etienne Barnard and David Casasent. Shift invariance and the neocognitron. *Neural Networks*, 3 (4):403–410, 1990.

Erik J Bekkers. B-spline cnns on lie groups. *arXiv preprint arXiv:1909.12057*, 2019.

Erik J Bekkers, Maxime W Lafarge, Mitko Veta, Koen AJ Eppenhof, Josien PW Pluim, and Remco Duits. Roto-translation covariant convolutional networks for medical image analysis. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pp. 440–448. Springer, 2018.

Johann Brehmer, Sönke Behrends, Pim de Haan, and Taco Cohen. Does equivariance matter at scale? *arXiv preprint arXiv:2410.23179*, 2024.

Anadi Chaman and Ivan Dokmanić. Truly shift-invariant convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3773–3783, 2021.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2990–2999, New York, New York, USA, 20–22 Jun 2016. PMLR.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 113–123, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.

Alhussein Fawzi, Horst Samulowitz, Deepak Turaga, and Pascal Frossard. Adaptive data augmentation for image classification. In *2016 IEEE international conference on image processing (ICIP)*, pp. 3688–3692. Ieee, 2016.

Rafael C Gonzalez. *Digital image processing*. Pearson education india, 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Alex Hernández-García and Peter König. Data augmentation instead of explicit regularization. *arXiv preprint arXiv:1806.03852*, 2018.

Daniel Ho, Eric Liang, Xi Chen, Ion Stoica, and Pieter Abbeel. Population based augmentation: Efficient learning of augmentation policy schedules. In *International conference on machine learning*, pp. 2731–2741. PMLR, 2019.

Emiel Hoogeboom, Jorn WT Peters, Taco S Cohen, and Max Welling. Hexaconv. *arXiv preprint arXiv:1803.02108*, 2018.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

Sékou-Oumar Kaba, Arnab Kumar Mondal, Yan Zhang, Yoshua Bengio, and Siamak Ravanbakhsh. Equivariance with learned canonicalization functions. In *International Conference on Machine Learning*, pp. 15546–15566. PMLR, 2023.

Eric Kauderer-Abrams. Quantifying translation-invariance in convolutional neural networks. *arXiv preprint arXiv:1801.01450*, 2017.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Maxime W. Lafarge, Erik J. Bekkers, Josien P.W. Pluim, Remco Duits, and Mitko Veta. Roto-translation equivariant convolutional networks: Application to histopathology image analysis. *Medical Image Analysis*, 68:101849, 2021. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2020.101849. URL https://www.sciencedirect.com/science/article/pii/S1361841520302139.

Chen-Yu Lee, Patrick W Gallagher, and Zhuowen Tu. Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree. In *Artificial intelligence and statistics*, pp. 464–472. PMLR, 2016.

Diego Marcos, Michele Volpi, Nikos Komodakis, and Devis Tuia. Rotation equivariant vector field networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5048–5057, 2017.

Md Ashiqur Rahman and Raymond A Yeh. Truly scale-equivariant deep nets with fourier layers. *Advances in Neural Information Processing Systems*, 36:6092–6104, 2023.

Renan A Rojas-Gomez, Teck-Yian Lim, Alex Schwing, Minh Do, and Raymond A Yeh. Learnable polyphase sampling for shift invariant and equivariant convolutional networks. *Advances in Neural Information Processing Systems*, 35:35755–35768, 2022.

Dominik Scherer, Andreas Müller, and Sven Behnke. Evaluation of pooling operations in convolutional architectures for object recognition. In *International conference on artificial neural networks*, pp. 92–101. Springer, 2010.

Maurice Weiler and Gabriele Cesa. General e(2)-equivariant steerable cnns. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/45d6637b718d0f24a237069fe41b0db4-Paper.pdf.

Daniel E. Worrall, Stephan J. Garbin, Daniyar Turmukhambetov, and Gabriel J. Brostow. Harmonic networks: Deep translation and rotation equivariance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Matthew D Zeiler and Rob Fergus. Stochastic pooling for regularization of deep convolutional neural networks. *arXiv preprint arXiv:1301.3557*, 2013.

Richard Zhang. Making convolutional networks shift-invariant again. In *International conference on machine learning*, pp. 7324–7334. PMLR, 2019.

Xueyan Zou, Fanyi Xiao, Zhiding Yu, and Yong Jae Lee. Delving deeper into anti-aliasing in convnets. In *BMVC*, 2020.

Xueyan Zou, Fanyi Xiao, Zhiding Yu, Yuheng Li, and Yong Jae Lee. Delving deeper into anti-aliasing in convnets. *International Journal of Computer Vision*, 131(1):67–81, 2023.